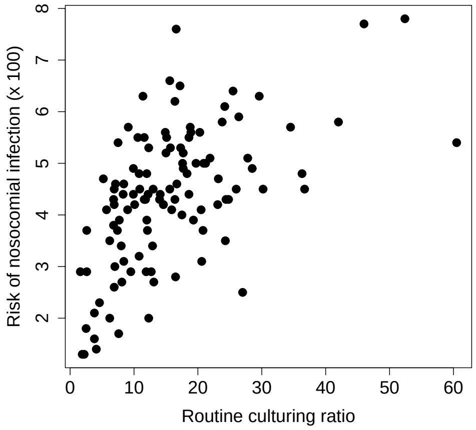
Correlation and the Analysis of Variance Approach to Simple Linear Regression

Biometry 755

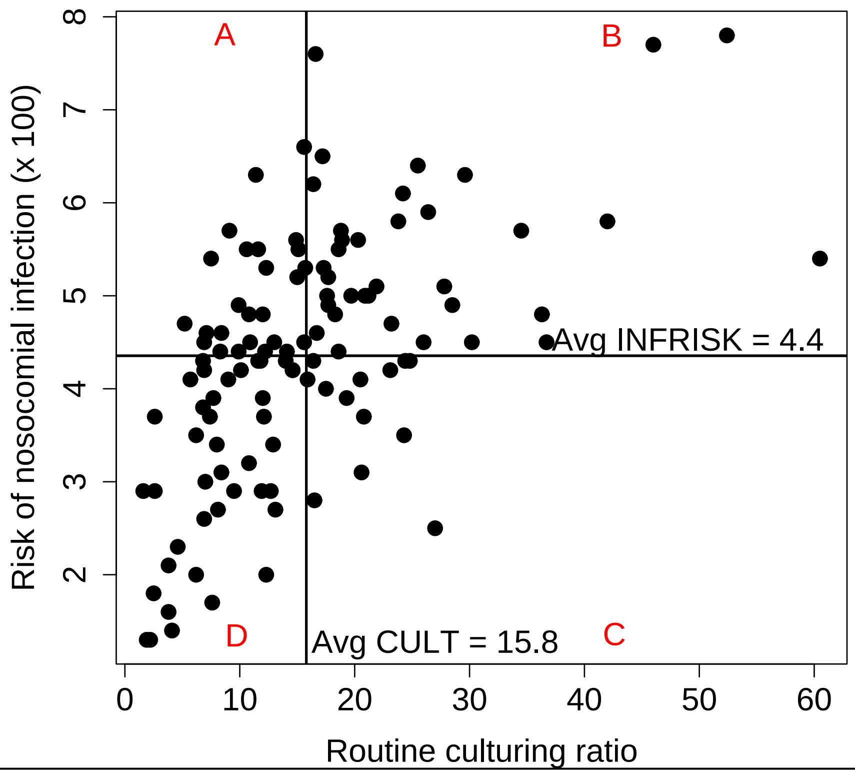
Spring 2009

Correlation review

Correlation quantifies the direction and strength of the linear association between two random variables. Consider the scatterplot of risk of nosocomial infection by routine culturing ratio. There appears to be a strong positively sloped linear relationship between the two variables. We would like a single index to quantify both features of this apparent relationship.



Quantifying linear association

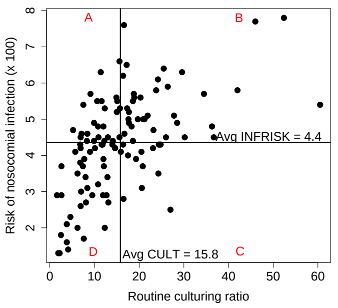


Quantifying linear association

Consider data points (x, y) in each of the four quadrants, A, B, C and D, formed by drawing a vertical line at the average culturing ratio value (X), and a horizontal line at the average value of nosocomial infection risk (Y).

Region	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})(y - \bar{y})$
A			
B			
C			
D			

The sample correlation coefficient



For points in quadrants A and C, $(x - \bar{x})(y - \bar{y})$ will be negative. For points in quadrants B and D, $(x - \bar{x})(y - \bar{y})$ will be positive. If a strong linear association exists, then the sum of this product across all data points,

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}),$$

will be dominated by either positive or negative terms.

The sample correlation coefficient (cont.)

(1) $$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

is an estimate of the *covariance* between X and Y, which measures the strength of their association. (Population covariance is denoted by σ_{xy}.)

It seems that Equation (1) is a good choice for assessing both direction and strength of linear association, but there is one drawback ... Equation (1) can be large because of the scale of measurement of the variables themselves, rather than the strength of a linear association. Therefore, we scale Equation (1) by dividing by estimates of the standard deviations of X and Y.

The sample correlation coefficient (cont.)

Recall that

$$\hat{\sigma}_x = s_x = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$$

and

$$\hat{\sigma}_y = s_y = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}}.$$

Then our ‘standardized’ index of linear association is

$$\frac{\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n-1}}{\sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}\sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}}} = \frac{s_{xy}}{s_x s_y}.$$

Definition of sample correlation coefficient

This leads to the following definition of the sample correlation coefficient, r. It is also known as the *Pearson correlation coefficient*.

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}.$$

- r’s range of values is -1 to 1.
- $r = 1 \Rightarrow$ observations lie on positively sloped line.
- $r = -1 \Rightarrow$ observations lie on negatively sloped line.
- r is a dimensionless measure.
- r measures the strength of the *linear* association.
- r tends to be close to zero if there is no linear association.

What does r estimate?

r is an index obtained from a sample of n observations and is an estimator for an unknown population parameter. The parameter is called the *population correlation coefficient*, and is defined as

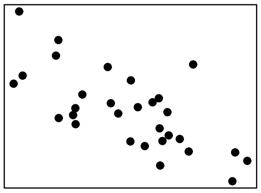
$$\rho = \frac{Cov(x,y)}{\sqrt{Var(x)Var(y)}} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}.$$

In other words,

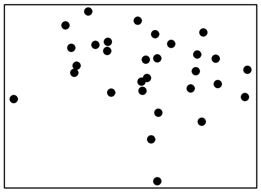
$$r = \hat{\rho}.$$

Picturing ρ and r

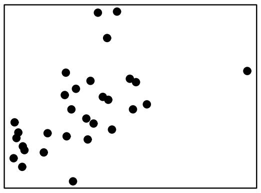
Each graph depicts a sample of 30 data points, (x, y), drawn from a population with the specified value of ρ. r is calculated based on the 30 data points.



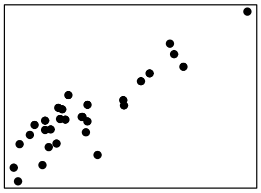
rho = −0.6 ; r = −0.691



rho = −0.05 ; r = −0.201



rho = 0.4 ; r = 0.556



rho = 0.9 ; r = 0.892

Inference about ρ

When r is non-zero, does that imply that ρ is non-zero? Not necessarily. We must have a method that accounts for the sampling variability in order to make rigorous inference about whether ρ is different from zero.

We use the following hypothesis testing procedure.

$$H_0 : \rho = 0 \text{ versus } H_A : \rho \neq 0.$$

The test statistic is

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

where r is the sample correlation coefficient and $t \sim t_{n-2}$ under H_0.

Correlation analysis in SAS

SAS's PROC CORR computes the sample correlation, r, and conducts a two-sided $\alpha = 0.05$-level test of the null hypothesis $H_0 : \rho = 0$.

```
proc corr data = one;
    var infrisk cult;
run;
```

PROC CORR output

```
 Pearson Correlation Coefficients, N = 113
         Prob > |r| under H0: Rho=0
                        INFRISK           CULT

INFRISK                 1.00000        0.55916
RISK OF INFECTION                       <.0001

CULT                    0.55916        1.00000
CULTURE RATIO            <.0001
```

We conclude, at $\alpha = 0.05$, that the true correlation between risk of nosocomial infection and routine culturing ratio differs significantly from zero.

CIs for correlation coefficients in SAS

Use the ‘FISHER’ PROC option in PROC CORR to obtain a 95% CI for the correlation coefficient. The CI is constructed based on ‘Fisher’s z transformation’ (see Rosner, Chapter 11).

```
proc corr data = one fisher;
    var infrisk cult;
run;
```

FISHER PROC option output

```
   Pearson Correlation Statistics
   (Fisher's z Transformation)

            With
Variable    Variable       95% Confidence Limits

INFRISK     CULT            0.415497      0.67288
```

Specifying other null values

Use the ‘RHO0’ FISHER option in PROC CORR to test the null hypothesis $H_0 : \rho = \rho_0$ where $\rho_0 \neq 0$.

```
proc corr data = one fisher (rho0 = 0.3);
    var infrisk cult;
run;
```

RHO0 option output

```
        Pearson Correlation Statistics
        (Fisher's z Transformation)

          With                               ------H0:Rho=Rho0-----
Variable  Variable  95% Confidence Limits   Rho0     p Value

INFRISK   CULT      0.415497      0.672880  0.30000      0.0008
```

Final comments about correlation in SAS

- SAS does not conduct a hypothesis test of $H_0 : \rho_1 = \rho_2$ (See Rosner, Section 11.11.4)
- Use Spearman's rank-order correlation coefficient when both variables are continuous but at least one is not normally distributed. (See Rosner, Section 12.6)
- Use Kendall's tau b correlation coefficient when at least one variable is ordinal. Here, we assume there exists an latent (unobserved) continuous variable underlying the ordinal variable.

Other correlation coefficients in SAS

```
proc corr data = one spearman;
    var infrisk cult;
run;
```

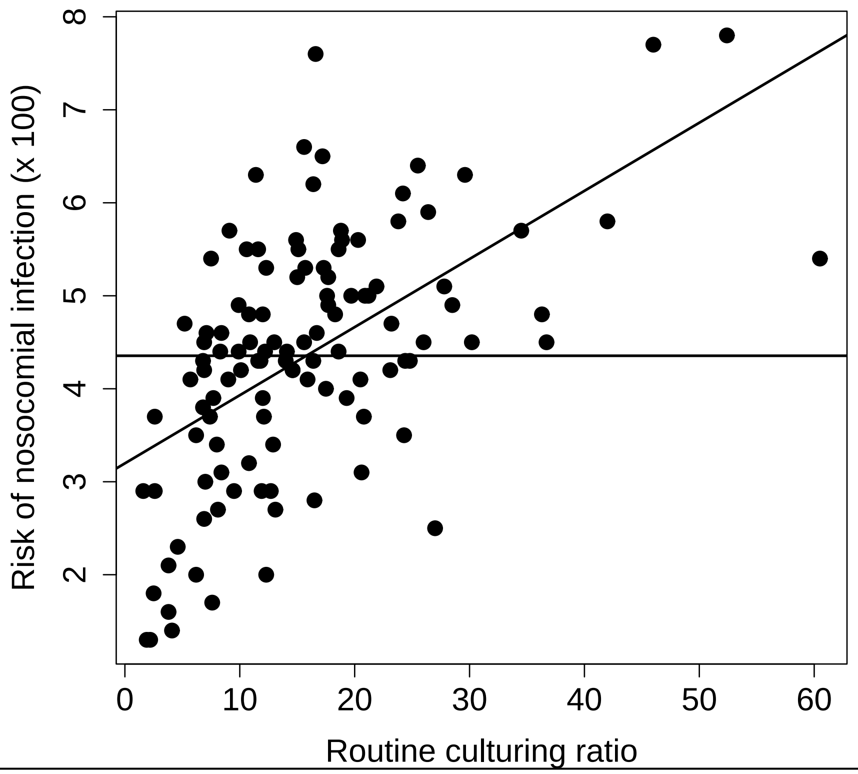
```
 Spearman Correlation Coefficients, N = 113
         Prob > |r| under H0: Rho=0

                        INFRISK          CULT

INFRISK                 1.00000       0.56036
RISK OF INFECTION                      <.0001

CULT                    0.56036       1.00000
CULTURE RATIO            <.0001
```

The analysis of variance approach to SLR

Our approach to SLR has been to find the straight line that best describes the linear relationship between X and Y. Another way of looking at this problem is as follows:

1. There is a certain amount of variability in the dependent variable, Y.
2. If we believe there is a linear association between X and Y, then this association accounts for some proportion of the observed variability in Y.
3. The best line is the one that ‘explains’ the greatest proportion of the total variability in Y.

The ANOVA approach to SLR (cont.)



Summarizing sources of variability in SLR

Given y_i, $\hat{y}_i$, and $\bar{y}$, the following statements are true*:

- $\sum_{i=1}^{n}(y_i - \bar{y})^2$ estimates the total variability in the set of y_is.
- $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ estimates the variability in the set of y_is explained by the regression of Y on X.
- $\sum_{i=1}^{n}(y_i - \hat{y})^2$ estimates the unexplained (*residual*) variability in the set of y_is.

* up to a constant of proportionality

Sources of variability in SLR (cont.)

This leads to the following fundamental principle of regression:

total variation $=$ variation due to regression $+$ residual variation.

That is:

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

or

$$\text{SSY}=\text{SSR}+\text{SSE},$$

where ‘SS’ stands for ‘sum of squares’, and ‘Y’, ‘R’ and ‘E’ are, respectively, ‘Y’, ’regression’ and ‘error’.

Sources of variability in SLR (cont.)

By simple algebra, we have

$$\text{SSR}=\text{SSY}-\text{SSE}.$$

Recall that our goal is to explain as much of the total variability in Y by the regression of Y on X. Therefore, we want the ratio

$$\frac{\text{SSR}}{\text{SSY}}=\frac{\text{SSY}-\text{SSE}}{\text{SSY}} \tag{2}$$

to be as large as possible. The closer the ratio in (2) is to 1, the better the regression at explaining Y’s variability.

Relating r to SLR

It can be shown that the square of the sample correlation coefficient is

$$r^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

or, as we've just seen,

$$r^2 = \frac{\text{SSY} - \text{SSE}}{\text{SSY}} = \frac{\text{SSR}}{\text{SSY}}.$$

Relating r to SLR (cont.)

In words, the square of the sample correlation coefficient is equivalent to the ratio of SSR to SSY. We therefore conveniently refer to SSR/SSY as R^2. We note the following properties of R^2.

1. R^2 is bounded between 0 and 1.
2. If $R^2 = 1$, then all of the variation in Y is explained by the regression.
3. If $R^2 = 0$, then none of the variation in Y is explained by the regression.

The ANOVA table for SLR

The results of a SLR can be conveniently summarized in an ANOVA table

Source	df	SS	MS	F
Regression	1	SSR = SSY - SSE	MSR $= \frac{\text{SSR}}{1}$	$\frac{\text{MSR}}{\text{MSE}}$
Residual	$n-2$	SSE	MSE $= \frac{\text{SSE}}{n-2}$	
Total	$n-1$	SSY		

where MSR stands for ‘mean square regression’ and MSE stands for ‘mean square error’.

The ANOVA table for the SENIC example

```
proc reg data = one;
    model infrisk = cult;
run;
```

```
                       Analysis of Variance
                             Sum of           Mean
Source               DF     Squares         Square    F Value    Pr > F

Model                 1    62.96314       62.96314      50.49    <.0001
Error               111   138.41668        1.24700
Corrected Total     112   201.37982

Root MSE    1.11669    R-Square     0.3127
```

ANOVA table for the SENIC example (cont.)

```
                            Analysis of Variance
                                   Sum of           Mean
Source                   DF       Squares         Square    F Value    Pr > F

Model                     1      62.96314       62.96314      50.49    <.0001
Error                   111     138.41668        1.24700
Corrected Total         112     201.37982

Root MSE    1.11669    R-Square     0.3127
```

SSR $\doteq 63.0$ SSE $\doteq 138.4$ SSY $\doteq 201.4$

$\hat{\sigma}^2 = s^2 =$ MSE $= \frac{\text{SSE}}{n-2} \doteq 1.2 \Rightarrow \hat{\sigma} = s \doteq 1.1$

$R^2 \doteq 0.31 \Rightarrow r \doteq 0.56$ ($r = \text{sign}(\hat{\beta}_1)\sqrt{r^2}$).

Approximately 31% of the total variability in risk of nosocomial infection is explained by its linear association with routine culturing ratio.

Overall test for significance of the regression

The value in the "F" column is the test statistic for the overall test for the regression. The general form of the test is as follows:

H_0: No linear association between X and Y
H_A: There is a linear association between X and Y

In SLR, testing for significance of the regression is equivalent to testing for significance of the slope. Therefore, we can restate the null and alternative hypotheses as

$H_0 : \beta_1 = 0$
$H_A : \beta_1 \neq 0$

Overall test (cont.)

Source	df	SS	MS	F
Regression	1	SSR = SSY - SSE	MSR $= \frac{\text{SSR}}{1}$	$\frac{\text{MSR}}{\text{MSE}}$
Residual	$n-2$	SSE	MSE $= \frac{\text{SSE}}{n-2}$	
Total	$n-1$	SSY		

$F =$ MSR$/$MSE will be small (close to zero) when the regression of Y on X fails to explain a meaningful proportion of Y's variability (i.e. when the null hypothesis is true). Alternatively, $F =$ MSR$/$MSE will be large when the regression explains a meaningful proportion of Y's variability (i.e. when the alternative hypothesis is true).

Overall test (cont.)

The test of the null hypothesis that there is no linear association between X and Y against the alternative that there is a linear association between X and Y is summarized as follows:

Test statistic: $F = \frac{\text{MSR}}{\text{MSE}} = \frac{\text{SSR}_{/1}}{\text{SSE}_{/(n-2)}}$.

Distribution under H_0: $F \sim F_{1,n-2}$.

P-value: p-value $= \Pr\left(F > \frac{\text{MSR}}{\text{MSE}}\right)$ where $F \sim F_{1,n-2}$. (Note that this is a one-sided test.)

Conclusion: Same as conclusion for test on the slope since, in SLR, the overall test is the same as the test on the slope.

SENIC example

```
                          Analysis of Variance
                                Sum of           Mean
Source                DF       Squares         Square    F Value    Pr > F

Model                  1      62.96314       62.96314      50.49    <.0001
Error                111     138.41668        1.24700
Corrected Total      112     201.37982
```

$F = \text{MSR}/\text{MSE} \sim F_{1,111}$. From the SAS output, MSR/MSE = 63.0/1.2 = 50.5, and $\Pr(F > 50.5) < 0.0001$, where $F \sim F_{1,111}$. Therefore, at $\alpha = 0.05$, we reject H_0 and conclude that there is a significant linear association between risk of nosocomial infection and routine culturing ratio.

Equivalence of F and t test in SLR

As a final note, we stated on Slide 30 that, in SLR, the overall test for significance of the regression is equivalent to the test for significance of the slope. In fact, it can be shown that

$$F_{1,\nu,1-\alpha} = t^2_{\nu,1-\alpha/2} = t^2_{\nu,\alpha/2}.$$

Equivalence of F and t test (cont.)

This can be verified from the PROC REG SAS output. The test-statistic for the test of slope is 7.11, the test statistic for significance of the regression is 50.49, and (7.11)^2 = 50.5521.

```
                                Sum of           Mean
Source              DF         Squares         Square    F Value    Pr > F

Model                1        62.96314       62.96314      50.49    <.0001
Error              111       138.41668        1.24700
Corrected Total    112       201.37982

                      Parameter Estimates
                     Parameter       Standard
Variable      DF      Estimate          Error    t Value    Pr > |t|

Intercept     1        3.19790        0.19377      16.50      <.0001
CULT          1        0.07326        0.01031       7.11      <.0001
```